Running head: Subjective Probability Distribution Width

Wide of the Mark: Evidence on the Underlying Causes of Overprecision in Judgment

Don A. Moore
University of California, Berkeley

Ashli B. Carter
Columbia University

Heather H. J. Yang
Massachusetts Institute of Technology

Abstract
Overprecision is the most robust and least understood form of overconfidence. In an attempt to elucidate the underlying causes of overprecision in judgment, the present paper offers a new approach--examining people's beliefs about the likelihood of chance events drawn from known probability distributions. This approach allows us to test the assumption that low hit rates inside subjective confidence intervals arise because those confidence intervals are drawn too narrowly. In fact, subjective probability distributions are systematically too wide, or insufficiently precise. This result raises profound questions for the study of overconfidence.

Draft of October 14, 2015

In press at *Organizational Behavior and Human Decision Processes*

Author note
Materials and data: http://learnmoore.org/BDE/. Email: don.moore@alumni.carleton.edu.

Wide of the Mark:

Evidence on the Underlying Causes of Overprecision in Judgment

Overprecision is the excessive faith that one's beliefs are correct. It is simultaneously the most robust and the least understood form of overconfidence (Moore, Tenney, & Haran, 2015). The typical overprecision study asks people questions with quantitative answers (e.g., "How much does Barack Obama weigh?") and asks them to estimate 90% confidence intervals around these answers. However, these 90% confidence intervals routinely contain the correct answer less than 50% of the time (Alpert & Raiffa, 1982). This overprecision effect is one of the most dramatic and impressive in the decision making literature, and has been replicated in many paradigms and populations (Block & Harper, 1991; Harvey, 1997; Mamassian, 2008).

However, one of the impediments to the study of overprecision has been the difficulty specifying the relevant knowledge that individuals possess and, consequently, whether they use that information effectively. Most prior research approaches have not made it easy to compare research participants' beliefs with the normatively correct beliefs at the level of the individual question (see Lawrence, Goodwin, O'Connor, & Önkal, 2006). Instead, when researchers observe that hit rates inside 90% confidence intervals are below 90%, they quite sensibly assume that this is because subjects are underestimating the uncertainty around their beliefs across the a set of questions (Tversky & Kahneman, 1974). However, without being able to compare how sure someone *is* with how sure they *ought to be* of something in particular, we cannot know whether judgmental overprecision is, in fact, always due to overly narrow subjective probability distributions.

This limitation imposes several problematic constraints. For one thing, it obscures the cause of overprecision because it cannot tell us whether or when confidence intervals have such low hit rates because they are too narrow or because they are centered on the wrong estimate. Many researchers write about overprecision as if it occurs because people have overly narrow subjective probability distributions (Harvey, 1997; Jain, Mukherjee, Bearden, & Gaba, 2013; Mannes & Moore, 2013; Soll & Klayman, 2004). Others posit that overprecision is a consequence of people relying on available, but potentially biased, information. For example, Juslin, Winman, and Hansson (2007) characterized people as "naïve intuitive statisticians." The naïveté of the intuitive statistician is the uncritical reliance on *sample* properties and mistaking them for *population* properties. Here, the underlying assumption is that people's beliefs are centered on sensible estimates gathered from experience, but that individuals fail to appreciate the fact that their small samples underestimates error variance—a mistake that leads to overly precise beliefs.

Hit rates below 90% appear as *prima facie* evidence that 90% confidence intervals are drawn too narrowly relative to the individual's own error distribution. Prior research has, however, relied on the untested assumption that, if only we could specify the error distribution (rather than inferring it post-hoc from observed error rates), we would observe that individuals' self-reported confidence intervals are too narrow. In this paper, we test this assumption. We can specify what participants *should* believe for a full range of outcomes and whether subjective probability distributions are in fact too narrow relative to this benchmark. Using a novel experimental paradigm, this paper questions whether low hit rates necessarily imply too much certainty in a particular belief as measured by confidence interval width. Furthermore, our

results sheds light onto the process that people use to make judgments in the face of uncertainty and provides evidence against some common assumptions in overconfidence research.

Explanations for Overprecision

The results will inform three of the most prominent explanations for overprecision in judgment: anchoring, conversational norms, and naïve intuitive statistics. The anchoring explanation holds that overprecision is the result of people first making some best estimate and the adjusting insufficiently from it (Block & Harper, 1991; Plous, 1995). If this explanation is right, then helping set the anchor by eliciting a best guess should lead to subject probability distributions centered even more tightly around the best-guess judgment. We do not find that it does.

The conversational norms explanation holds that people express overprecision because they are trying to provide informative judgments, even when that comes at the expense of accuracy. Imagine I ask my friend for the location of Stanford University. If my friend tells me it is in the city of Palo Alto, that would be informative but inaccurate, given that the University is, in fact, in the municipality of Stanford, California (next to the larger city of Palo Alto). However, my friend's response is more useful than if she would have said that Stanford is somewhere in northern California, a response that would have been accurate at the expense of being informative. Indeed, many people express a preference to get informative over accurate advice (Yaniv & Foster, 1995). For this to be able to account for overprecision, it must arise from overly precise subjective distributions. That is not what we observe in our data.

The naïve intuitive statistician argument holds that subjective error distributions are smaller than actual error distributions because our minds take a small sample of relevant facts. Our minds are limited to thinking of about 7 (plus or minus 2) facts at once (Juslin et al., 2007).

This small sample will have a smaller variance than the actual population of relevant facts, leading people to underestimate the uncertainty around their knowledge. This explanation also posits that subjective probability distributions are overly narrow, and it applies best to epistemic uncertainties that arise due to the imperfections in our own knowledge. It does a poor job explaining why we observe underprecision in subjective probability distributions, regardless of whether uncertainties are framed as either epistemic or aleatory.

Overview of the Studies

Experiment 1 tests the traditional method of eliciting confidence intervals against our new approach. Experiment 2 attempts to reconcile results from our new approach with apparently contradictory conclusions in the research literature. For Experiment 2 and the remaining experiments, rather than ask for a confidence interval, we use the Subjective Probability Interval Elicitation (SPIES) measure introduced by Haran, Moore, & Morewedge (2010) whereby participants estimate the full probability distribution of outcomes—providing a probability estimate (from 0 to 100%) for the likelihood of each possible outcome.

In Experiment 3, we examine the degree to which novel results from our new approach are due to its lack of familiarity. In Experiment 4, we explore the degree to which the expression of uncertainty in our new paradigm is moderated by its conceptualization as uncertainty around a repeatable event with many different possible outcomes based upon rules of chance (aleatory uncertainty) or as lack of knowledge regarding a unique event with a particular outcome (epistemic uncertainty), (Fox & Ülkümen, 2011). We explore this possibility because past work shows that individuals express less certainty when making judgments of events that are unknown due to chance factors (aleatory frame) compared to events that are unknown due simply to lack

of knowledge (epistemic frame), (Fox & Ülkümen, 2011). Finally, Experiment 5 examines the robustness of our results using a behavioral measure of precision in judgment.

We report how we determined our sample size, all data exclusions (if any), all conditions, and all measures for all studies. Data and materials are available online: http://learnmoore.org/BDE/ .

EXPERIMENT 1

Method

In order to compare our new approach to traditional methods, we conducted a survey with three general knowledge questions of the sort that have consistently produced overprecision in prior research: Estimating the increase in the value of a stock, estimating someone's weight, and estimating someone's age (Gino & Moore, 2007). To these, we added two questions with known probability distributions, shown in Table 1 (estimating the final location of a jumping bean and estimating the winnings of a lottery).

Table 1. The five question topics used in Experiment 1.

Topic	Exact wording of question	Best answer
Apple Stock	Suppose you had invested $100 in the stock of Apple Computer Inc. (of Cupertino, CA) on January 1, 1992 and you didn't sell any stock. What is your guess about how much that investment would have been worth on October 1, 2013?	$3533.38[a]
Weight	Enter your estimate of the person's weight [from a photograph]	147.8 lbs[b]
Age	How old is the person in the picture above? [photograph]	41 years[c]
Jumping Bean	Suppose a jumping bean is lying on a sloped sidewalk. Each jump has a 75% chance of moving the bean 1 inch to the left and a 25% chance of moving the bean 1 inch to the right. How many inches away from its starting point will the jumping bean be after 600 jumps?	300 inches to the left[d]
Lottery	Suppose you are planning to participate in a lottery game. Each day there is a 60% chance you will win $1 and a 40% chance that you will lose $1. How much money will you end up with after 500 days?	$100[e]

Best answers were determined by [a] historical Apple share price data adjusting for dividends and splits, [b] actual weight of person in the photograph, [c] actual age of person in photograph, [d] most likely location after 600 jumps, and [e] most likely winnings after 500 days.

Design. All participants answered questions about each of the five topics, presented in random order. For the lottery topic questions, following Jain, Mukherjee, Bearden, and Gaba (2013), we tried to help our participants understand the lottery's random nature by providing a picture of 10 random paths of winnings that were possible for the first 150 days of the lottery. However, because we were concerned about participants anchoring their judgments on this sample of 10 random paths, we generated 20 such pictures (each with 10 different paths) and randomly presented each participant with one of them. We were also concerned that the figure's scale would provide an implicit possible range of possible winnings, so we generated two versions: ten of the pictures had a scale that ran to $100 (where the 10 paths were easy to distinguish), and the other ten had a scale that ran to $500, the theoretical maximum. The pictures presented to participants appear in this paper's online supplement. These precautionary variations did not end up producing any significant effects on our results, so we do not dwell on them.

For each of the five topics, we asked two questions that have been used in prior research on overprecision in judgment:

1) *90% confidence interval*: "Please give us two numbers: a 'lower bound' and an 'upper bound'. The 'lower bound' is a number so low that there is only a 5% probability that the right answer is less than that. Similarly, an 'upper bound' is a number so high that there is only a 5% probability the right answer is more than that. In other words, you should be 90% sure that the answer falls between the lower and upper bounds."
2) An *item-confidence judgment*, which had two parts:

a. A “best estimate” of the right answer,
b. A confidence question: “How confident are you that your answer is within 5% of the right answer?”

In the Apple stock, weight, and age conditions, participants were given a blank space to type in both their upper and lower bound estimates for the 90% confidence interval elicitation with the correct unit of measurement (dollars, pounds, and years) provided. For the jumping bean and lottery conditions, participants responded using a sliding bar for both the lower and upper bound that they could drag to indicate their estimate. In the center of the sliding bar was always zero and at each end were the maximum values (600 inches to the left or 600 inches to the right for the jumping bean condition and $500 loss and $500 gain in the lottery condition).

For the item-confidence elicitations, participants in all conditions were provided with a blank space to indicate their best estimate of the right answer (i.e., How many inches away from its starting point will the jumping bean be after 600 jumps? To the left of where it started or to the right?). Participants then used a sliding bar to indicate their confidence that their best estimate was within 5% of the correct answer. The sliding bar ranged 0% to 100%. Participants answered these two questions (90% confidence interval and item-confidence judgment) in randomly-determined order.

Participants. We opened the survey to 200 workers on Amazon’s Mechanical Turk (Buhrmester, Kwang, & Gosling, 2011). We selected that sample size arbitrarily based on a guess about the appropriate sample size, but did so before collecting any results. Participants each received $.25. After completing the consent form, participants had to pass an attention check before continuing to the experiment. In all the experiments we report, those who failed the attention check were rejected before they saw any of the key dependent measures and so are not

included in subject counts. The attention check and the rest of the survey appear online in the supplemental materials.

We expected to replicate prior findings of overprecision using traditional measures such as confidence interval hit rate for all five topic questions. More specifically, we expected participants' 90% confidence intervals to contain the correct answer significantly less than 90% of the time for all five topic questions. In line with reasoning that overprecision is caused by overly precise subjective probability distributions, we also expected to find that 90% confidence intervals were narrower than the true probability distributions for the lottery and jumping bean questions.

Results and Discussion

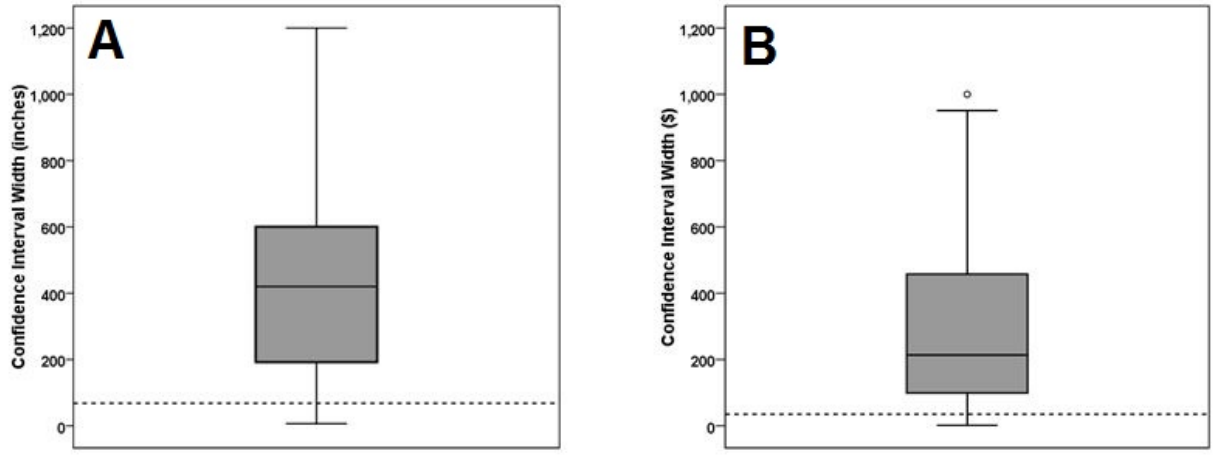
Using traditional measures for question format, we observe systematic overprecision in the results. Hit rates inside 90% confidence intervals are only 58%. This difference is significant for all five questions, $t\ (199) > 4$, $p < .001$. And for the item confidence question, while participants are, on average, 47% confident that they have guessed the right answers within 5%, they are only this close 25% of the time. This difference is significant for four of the five questions, $t\ (199) > 9$, $p < .001$. So using traditional measures of precision in judgment, our participants generally look over-precise. See Table 2.

Table 2. 90% confidence intervals and item-confidence judgments (Experiment 1).

Question	90% CI hit rate	Mean CI width (and SD)	Optimal CI width	Average Item-confidence	Item-Confidence Hit rate
Apple Stock	27%	$2,422,231 (12,882,834)	unknown	32%	0.5%
Weight	65%	41.2 pounds (29.3)	unknown	56%	30%
Age	78%	16 years (10)	unknown	51%	36%
Jumping Bean	51%	426 inches (258.6)	68 inches[a]	46%	11%
Lottery	70%	$289 (224)	$35[b]	50%	49%

Optimal confidence interval widths determined by [a] the span of inches where there is a 90% probability of the bean landing based upon the likelihood of the bean jumping to the left or right at each jump for 600 jumps (ranges from 266 inches to the left to 334 inches to the left) ; and [b] the span of dollar amounts where there is a 90% probability of winning based upon the likelihood of gaining or losing $1 each day after 500 days of the lottery game (ranges from $82.50 to $117.50).

Figure 1. Jumping Bean and lottery CI width distributions (Experiment 1)



Note: Optimal CI Width shown as dashed line

For the lottery and bean questions, where the true outcome probabilities are easy to identify, we have an additional measure of the precision of judgment: Are reported confidence intervals too narrow? The answer is an emphatic and surprising no. By this measure, our respondents are catastrophically *under*precise: their intervals are far too wide. For the jumping bean, the optimal 90% confidence interval is centered on a move of 300 inches to the left from the origin and is 68 inches wide. On average, participants' confidence intervals were 426 inches wide—over 6 times as wide as they should have been. Figure 1 shows the full distribution of reported confidence interval widths, and demonstrates that the excessively wide average reported confidence interval was not the result of a few misguided participants. Most participants did not even come close to the correct width of 68 inches. Hit rates are still below 90% (a traditional

indicator of *over*precision), however, this is because they are centered on the wrong value. The average "best guess" was 230 inches left (*SD* = 237 inches).

For the lottery, the optimal 90% confidence interval is centered on a $100 gain and is $35 wide. On average, participants' confidence intervals were $289 wide, an amazing 8 times as wide as they should have been. Figure 1 shows that the large average confidence interval width was not driven by outliers—most participants reported far too wide confidence interval widths that were far from the correct width of $35.[1] As with the lottery, their hit rates are still below 90% because these intervals are centered on the wrong value. The average "best guess" was $159 (*SD* = $117). These findings suggest that while 90% confidence interval hit rates are consistently low for all questions, this low hit rate is not necessarily caused by overly narrow confidence intervals. In fact, when we know how wide confidence intervals *should be*, in this case for the jumping bean and lottery questions, participants are actually *under*precise.

EXPERIMENT 2

Given the surprising results of Experiment 1, we next sought to replicate both overprecision (relative to hit-rate accuracy) and underprecision (relative to actual probability distribution), testing the robustness of Experiment 1's results in a context that we hoped would be more comprehensible and familiar than the ones we used in Experiment 1. Experiment 2 compared individuals' beliefs about outcomes in domains of increasing familiarity in an attempt to reconcile prior findings of overprecision in subjective probability distributions (Moore &

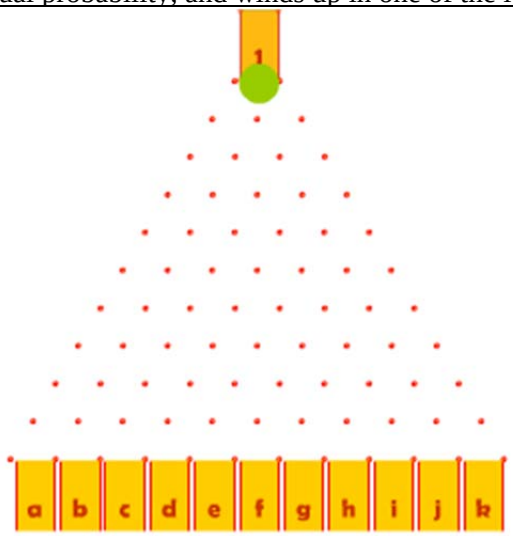
[1] It is noteworthy that these results appear to contradict the conclusions of Jain et al. (2013), who use a similar paradigm but who report overprecision in their results. The simple explanation is that Jain et al. elected to drop from their analysis any respondent who provided a confidence interval that was, in the researchers' opinion, too large (Kriti Jain, personal communication, July 3, 2013). Confidence intervals among the remaining participants were, of course, too small.

Healy, 2008) with the underprecision we found in Experiment 1. In Experiment 2, participants either gave estimates of performance on a trivia quiz or estimates about the outcomes of a chance device. We wanted to expand our investigation of the domains in which underprecision occurred, and compare our results against those from SPIES elicitations used to measure the full subjective probability distribution. In particular, we included a condition whose paradigm mirrors the SPIES elicitation used by Moore and Healy (2008). They gave their participants 10-item trivia quizzes and then asked them to estimate the chances that they (and others) obtained each of the 11 possible scores (from 0 to 10). Moore and Healy found that participants reported probability distributions that were too narrow. In particular, participants reported being excessively sure that they knew how others had scored on the quizzes. Experiment 2 asked participants to estimate the distribution in others' scores on a quiz, allowing us to test whether the results would replicate those of Moore and Healy, or whether they would replicate the underprecision observed in Experiment 1.

Method

Design. Participants were randomly assigned to one of four experimental conditions. The first condition, with the context of judgment least familiar to participants, employed a chance device: a ball-drop machine (the Quincunx) in which a ball, dropped from top, bounced down over 10 rows of pegs (see Figure 2). Participants were told that at each peg, the ball could bounce to either the left or the right with equal probability. While it is reasonable to believe that some participants may have seen this type of ball-drop machine before, we believed that the Quincunx would be less familiar than the context of judgment in the other conditions described below.

Figure 2. The Quincunx: A green ball drops in from the slot on top. At each peg, it bounces either to the left or the right with equal probability, and winds up in one of the lettered bins at the bottom.



The other three conditions asked participants to estimate scores on a difficult trivia quiz consisting of ten difficult true/false items, a context presumably more familiar to them. The quiz items appear in the online materials. Pre-testing established that average scores on the quiz were about 50%. Participants were informed, "Each question has two options, and, on average, there is roughly a 50% chance of each answer choice being correct."

Within these three quiz conditions, there were two "other-quiz" conditions and one "self-quiz" condition. In the two "other-quiz" conditions, participants were asked to estimate the score of another person who was randomly selected from a large group that had taken the quiz.

- In the low-information version of the other-quiz condition, participants did not see any of the quiz items themselves.
- In the other-quiz high-information condition, participants took the quiz themselves before estimating another person's score. Participants in this condition did not receive any indication of their own score before estimating another person's.
- In the self-quiz high-information condition, participants first took the quiz and then estimated their own scores.

For all four conditions, participants had to indicate how likely they thought each outcome was. We asked for this full subjective probability distribution rather than an item-confidence question in the hopes that it would induce deeper thought and more accurate responding. In the SPIES elicitation, participants estimated the probability of each possible outcome (e.g., bin in which the ball could drop) ranging from 0 to 100%. Each slider did not have an initial starting point (in order to prevent anchoring) and produced a number only once the participant clicked somewhere on the slider between 0 to 100%. Participants had to click each slider at least once before advancing, even to indicate a probability of 0%. For the Quincunx condition, participants reported how likely it was that the ball would land in each of the 11 bins at the bottom of the machine. In the three quiz conditions, participants reported how likely each quiz score was. Each slider required a response between 0 and 100%, but participants were not forced to sum all the probabilities to 100%. When participants gave answers that totaled something other than 100%, we standardized them to sum to 100%. We did this by dividing each answer buy the sum total so that they reflected the original proportions that the participant gave but on a 0 to 100% scale.

We expected to replicate the overprecision found in Experiment 1 using more traditional measures in all four conditions. In other words, when we look at the bin or quiz score assigned the greatest probability of likelihood as a measure similar to item-confidence and compare it to the actual probability of that outcome, we expected participants to believe too strongly in its likelihood. We can also compute the variance of an individual's reported subjective probability distribution as we would the variance of any distribution, by multiplying the mass at each point by the square of its distance from the distribution's mean. While we expected to replicate the underprecision found in Experiment 1 for this measure of variance, we hypothesized that as the

domain became more familiar, that the variance in reported subjective probability distributions would decrease. In particular, variance would be greatest for the Quincunx, followed by other-quiz low information, followed by other-quiz high information, followed by self-quiz high information.

Participants. We opened the survey to 400 MTurkers who participated in exchange for $.25 and a chance at a $25 prize. We selected that sample size based upon a guess of the appropriate sample size, before collecting any results. Participants who passed two attention checks were informed that their chances to win the prize increased with the accuracy of their responses, according to the quadratic scoring rule (Selten, 1998).

Results and Discussion

We first looked at the bin or quiz score assigned the greatest probability of likelihood as a measure similar to item-confidence and compared it to the actual probability of that outcome. We find, as in Experiment 1, that participants believe too strongly in the outcome they rate as most likely. Participants on average believe that the probability of the ball landing in their favored bin (in the Quincunx condition) or of a particular score (in the three quiz conditions) is 19.5%. The actual likelihood of the selected outcome is about 10.3%. This difference is significant for all four conditions, all ts > 8, $p < .001$.

However, as expected, there was a significant effect of the experimental manipulation on the variance in reported subjective probability distributions, F (1, 398) = 25.6, $p < .001$. The means (and SDs) appear in Table 3.

Table 3. The mean confidence assigned to the bin/score rated most likely, the actual probability of that being the right answer, the variance of reported probability distributions, and the variance of the normative probability distribution across four experimental conditions (Experiment 2).

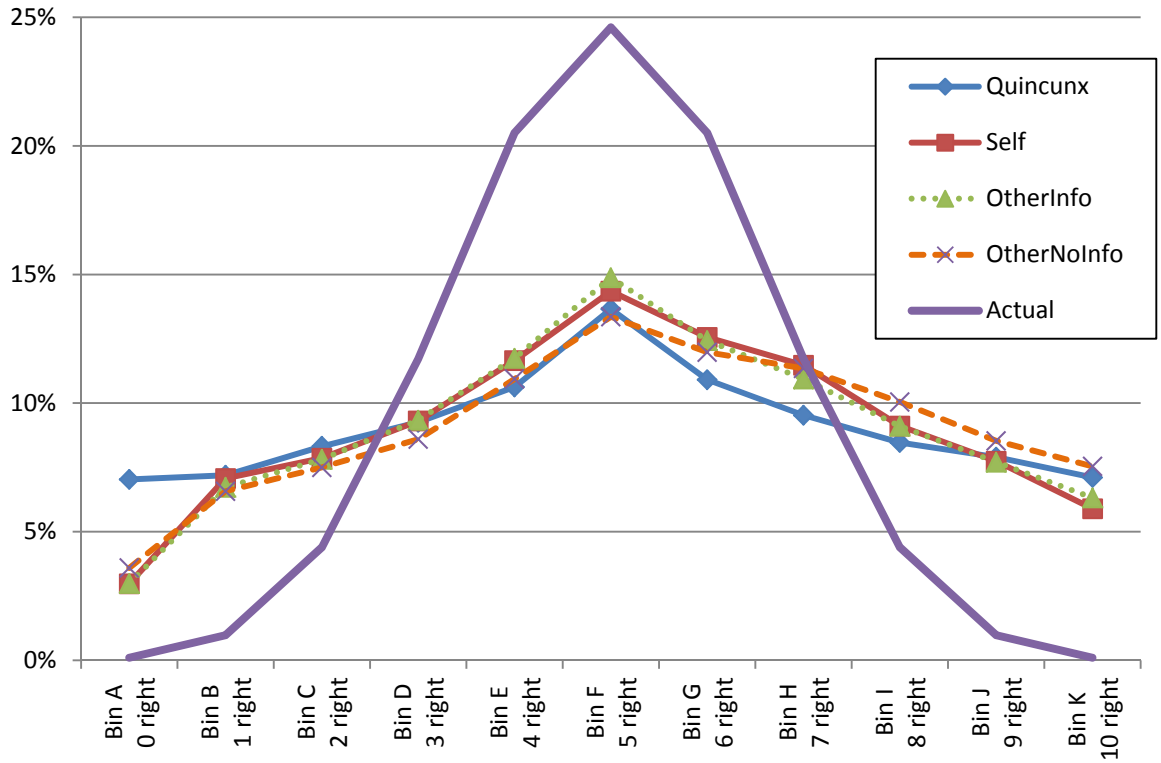
Condition (From least to most familiar)	SPIES peak probability	Peak bin/score actual probability	Mean variance of reported subjective probability distribution (and SDs of those means)	Variance of normative probability distribution [a]
Quincunx	17%	13.4%	9.33 (3.39)	3.5
Other-quiz low information	20%	11.3%	7.02 (2.64)	3.5
Other-quiz high information	20%	12.1%	6.95 (2.16)	3.5
Self-quiz	21%	11.3%	6.64 (2.77)	3.5

[a] Variance of the normative probability distribution determined by multiplying the mass at each point (probability of each given outcome) by the square of its distance from the distribution's mean. The probability for each outcome is as follows: Bin A/Score of 0 out of 10 - 0.1%; Bin B/Score of 1 out of 10 - 0.1%; Bin C/Score of 2 out of 10 - 4.4%; Bin D/Score of 3 out of 10 - 11.7%; Bin E/Score of 4 out of 10 - 20.5%; Bin F/Score of 5 out of 10 - 24.6%; Bin G/Score of 6 out of 10 - 20.5%; Bin H/Score of 7 out of 10 - 11.7%; Bin I/Score of 8 out of 10 - 4.4%; Bin J/Score of 9 out of 10 - 0.1%; Bin K/Score of 10 out of 10 - 0.1%. The formula for the distribution mean is as follows: ((0.1*1 + 0.1*2 + 4.4*3 + 11.7*4 + 20.5*5 + 24.6*6 + 20.5*7 + 11.7*8 + 4.4*9 + 0.1*10 + 0.1*11)/100) = 6. The resulting formula for the variance of the normative probability distribution is ((0.1*(0-6)^2 + 0.1*(1-6)^2 + 4.4*(2-6)^2 + 11.7*(3-6)^2 + 20.5*(4-6)^2 + 24.6*(5-6)^2 + 20.5*(6-6)^2 + 11.7*(7-6)^2 + 4.4*(8-6)^2 + 0.1*(9-6)^2 + 0.1*(10-6)^2)/100) = 3.5.

The main effect of the experimental manipulation reveals that, as expected, that participants reported more precise beliefs about people's scores on a quiz than about a ball in the Quincunx, as measured by variance of subjective probability distribution. A one-way ANOVA shows that participants reported the greatest variance in the Quincunx condition (the least familiar) compared to all other conditions, all t values > 5.74, p values < .001. While the variance reported in the quiz conditions did not significantly differ from one another, all t values < .97, p values > .34, we do see a slight trend whereby reported variance increased as the quiz became less familiar as predicted. Importantly, variances of reported subjective probability distributions in all four conditions are substantially greater than the normative benchmark of 3.5, all t values > 11, p values < .001. In sum, the degree of underprecision is moderated, as

expected, by our experimental manipulation whereby underprecision increased as familiarity with the context of judgment decreased. However, Figure 3 shows that all conditions continue to exhibit underprecision regardless of context of judgment using our new approach, whereby we can specify what individuals' subjective probability distribution *ought* to be.

Figure 3. Reported probability distributions (SPIES) in the different conditions of Experiment 3, along with the steeper line for the actual normative binomial probability distribution.



Results from Experiment 2 suggest that the underprecision observed in Experiment 1's 90% confidence intervals was not purely attributable to that elicitation format. Eliciting the full probability distribution again produced underprecision, which was worst with a chance device, the Quincunx. Importantly, we replicate overprecision for the SPIES peak probability measure that could not be explained by probability distributions that are too narrow. Instead, we see that across all four conditions, only 27% of participants assign the greatest probability to the most likely bin, Bin F, (in the Quincunx condition) or most likely quiz score, 5 out of 10.

In Experiment 2, underprecision decreased as participants became more familiar with the context of judgment. However, it is difficult to assess the degree to which participants' experience with other quizzes may have influenced their beliefs. Experiment 3 sought to address this concern by directly manipulating familiarity with the Quincunx. We suspected that greater familiarity with the Quincunx would improve accuracy and reduce underprecision.

EXPERIMENT 3

The more unbiased experience people have, the better able they are to reason about the properties of uncertain facts, forecasts, and probability distributions (Fiedler, 2000). The design of Experiment 3 sought to vary how easy it was for participants to acquire exactly such unbiased experience with the probability distribution of ball-drops in the Quincunx machine. Note that the acquisition of this useful experience also allows us to address another potential concern about the imperfections in our participants' responses. There are some biases evident in decision making under uncertainty that reduce, change, or disappear when people get to experience that uncertainty first-hand rather than considering it in the abstract (Hertwig, Barron, Weber, & Erev, 2004; Hertwig & Erev, 2009). By providing participants with personal first-hand experience we sought to address questions about whether the overly broad probability distributions we observe are restricted to decisions from description.

Method

Design. Each participant saw one of three Quincunx machines: a still photograph as in Experiment 2 (see Figure 2), a slow machine, or a fast machine. The slow machine simulated a ball drop when participants clicked the slot. Participants were able to click to see a single ball drop through the pegs and land in bin as many times as they wanted. The fast machine dropped

balls as a rate of about thirty per second, and participants could watch as they accumulated in the bins. Participants in all three conditions could take as much time as they wanted to observe their Quincunx machine before they completed the survey, and for all machines, there was a 50% chance of the ball going to the left or right at each peg.

All participants made two predictions regarding the probability of where the ball would land in a future ball drop: item confidence and SPIES. The item-confidence (IC) question asked participants to identify the one bin in which they thought the ball was most likely to land and then to specify the probability it would land there, as in Experiment 1. The SPIES elicitation asked for the full probability distribution across bins, as in Experiment 2.

Participants. Because we were concerned about the possibility that our results could be influenced by inattentive online participants, we compared responses from MTurkers and undergraduates working in the experimental laboratory of a west coast U.S. university. Laboratory and academic schedules allowed us to obtain responses from 74 students, which we compared against an equal number of responses from MTurk participants. Participants received a small fixed payment and the chance to win one of several $40 prizes. Participants were rewarded for accurate predictions by earning lottery tickets for the $40 gift card prizes.

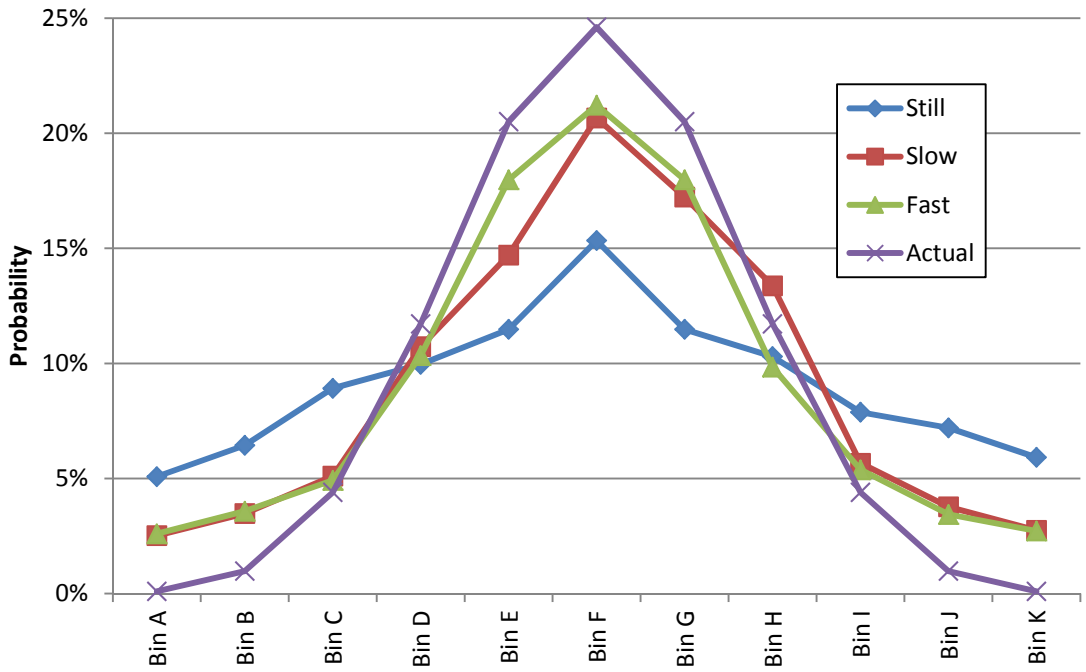
Results and Discussion

The traditional item-confidence measure again produces what appears to be overprecision. Participants estimated the likelihood that the ball would go into their chosen bin at 32% (SD = 21.4). In reality, the likelihood that the ball would go into that bin was 19%, a significant difference by paired t (147) = 7.59, $p < .001$. Again, we observe what appears to be overprecision using this standard measure. It is noteworthy that only 51% of participants chose the most likely bin, Bin F, for the item-confidence measure.

To analyze their SPIES reports, we began by focusing on the one bin that each participant rated as most likely. As with item-confidence, their confidence (M = 22.6%, SD = 9.9%) exceeded the actual probability (M = 19.5%, SD = 8.0%), paired t (147) = 3.30, p = .001. However, replicating the better calibration often observed with SPIES (Haran, Moore, & Morewedge, 2010), their overconfidence was substantially reduced relative to the item-confidence question, paired t (147) = 6.34, $p < .001$. The appearance of overprecision is easily exacerbated then by item-confidence elicitations, at least relative to SPIES. Additionally, a slight majority of participants—59%—assigned the actual most likely bin, Bin F, the highest probability for the SPIES elicitation as can be seen in Figure 4. (Participants who saw the still machine chose the correct bin 48% of the time, those who saw the slow machine 58%, and the fast machine 64%, averaging to 59%.)

Another way to measure precision, as introduced in Experiment 1 and 2, is to compare the variances of the Quincunx's actual probability distribution with participants' SPIES probability distributions. The actual variance (3.5) is significantly *smaller* than variance of participants' SPIES reports (6.21), paired t (147) = 10.2, $p < .001$. Again, by this measure people are underprecise, reporting subjective probability distributions that are too wide, as shown in Figure 4.

Figure 4. Reported probability distributions (SPIES) with varying familiarity (Experiment 3).



Furthermore, our independent variable (the degree of experience with the Quincunx) affected participants' calibration. A MANOVA examining the effect of our between-subjects manipulation of information on peak levels of confidence using both item-confidence and SPIES measures reveals that the more experience participants had with the Quincunx, the more precise their judgments became, F (2, 145) = 8.7, p = .001, as shown in Table 4. Experience increased the frequency with which participants correctly identified Bin F as most likely for both the item-confidence and SPIES measures, and it also increased their confidence that the ball would wind up in their chosen bin for both measures. See Table 4. The actual probability for the most likely bin, Bin F, is 24.6%.

Table 4. Results for experimental conditions (Experiment 3).

Information condition	IC confidence	IC actual prob. of chosen bin	IC pct choosing Bin F	SPIES peak confidence	SPIES actual prob. of chosen bin	Mean SPIES variance (and SDs)	SPIES choosing Bin F as most likely
Still photo	24.6%	15.5%	34%	18.5%	17.6%	8.3[a] (3.18)	48%
Slow machine	33.3%	19.9%	57%	24.2%	20.4%	5.6[b] (2.64)	60%
Fast machine	39.3%	22.2%	64%	25.3%	20.6%	4.7[b] (2.73)	64%

Different superscripts within SPIES variance column imply statistically significant difference.

Although the fact that their item-confidence judgments averaged 32%, while the maximum probability associated with any bin was 24.6% does imply that people believed too fervently that they knew the right answer, the peak of the subjective probability distribution flattened substantially when people thought about the full distribution on the SPIES question. In addition, we replicate the underprecision from Experiments 1 and 2: the variance in participants' SPIES distributions exceed the normatively correct value of 3.5 in all conditions. However, participants who received more information (especially the fast machine) came closer to that correct value.

None of these results are moderated by subject pool (college students in the lab vs. MTurk participants online). This fact should either give us either confidence in data collected online or concern about the diligence of our undergraduate participants.

EXPERIMENT 4

One striking feature of our known-uncertainty questions is that it is tempting to think about the uncertainty associated with them as aleatory rather than epistemic (Fox & Ülkümen, 2011). Aleatory uncertainty is best represented by a probability distribution of repeatable chance

outcomes. Under aleatory uncertainty, an event, if repeated under similar conditions, may have different outcomes due to unpredictable, chance causes. Stated differently, aleatory uncertainty stems from events that are fundamentally random, such as flips of a coin flip (Tannenbaum, Fox, & Ülkümen, 2014). By contrast, epistemic uncertainty is marked by a lack of knowledge of a unique unrepeatable event—knowledge that is, in principle, knowable. This type of uncertainty is due to missing information or expertise rather than stochastic causes. For example, uncertainty around whether you answered a single item on an exam correctly is epistemic uncertainty.

Because events can entail both aleatory and epistemic uncertainty, it is possible to experimentally manipulate the way a single event is framed or described. Past work shows that aleatory frames produce greater perceived uncertainty and wider subjective probability distributions perhaps because they draw attention to the random nature of an event while epistemic frames direct focus to the elements of an event that are knowable (Tannenbaum et al., 2014). Experiment 4 sought to manipulate this framing directly to see if epistemic framing would eliminate our findings of underprecision when measured by variance of subjective probability distributions. While we were unsure whether epistemic framing would eliminate our findings of underprecision, we did expect participants in this condition to be more precise in their item-confidence and SPIES estimates than those in the aleatory framing condition.

Method

Design. All participants saw the still picture of the Quincunx (See Figure 2) and were told about the ball-drop machine. To simulate a lack of knowledge of a unique event, in the epistemic condition, we told participants that a ball had already been dropped and to report their confidence of how likely it was that it had landed in each of the possible bins. Importantly, the

epistemic condition described a unique, unrepeatable event whose outcome was in principal knowable. To highlight this further, we also told participants that we would show them the actual outcome after they answered. In the aleatory condition, to mirror a probability distribution of repeatable chance outcomes, we asked participants to estimate the proportion of balls that would land in each of the possible bins if we dropped the ball 100 times. Under this framing, we describe the ball drop as a repeatable event. By asking participants to estimate the proportion of balls that would land in each of the bins after 100 drops, we highlight that each single ball drop event may have a different outcome due to chance.

All participants answered the item-confidence and the SPIES questions, in counterbalanced order.

Participants. Guided by a power analysis that assumed a small (Cohen's *d* of .3) between-subjects effect size and aimed for 80% power, we opened the survey on MTurk to 352 participants, offering $.20 and a chance at one of four $40 prizes. After dropping participants who failed to complete all the answers, we were left with 339 observations.

Results and Discussion

We expected those in the epistemic condition would be more likely to claim that they knew the answer. Indeed, those in the epistemic condition reported a mean item-confidence of 33% that they had correctly identified the right bin, compared with 25% in the aleatory condition. This difference is statistically significant in a 2 (epistemic vs. aleatory) X 2 (order: SPIES vs. item-confidence first) ANOVA, F (1, 335) = 15.4, $p < .001$. This represents a replication of the results of Tannenbaum et al. (2014) on epistemic/aleatory frames. There is no significant effect of order, F (1, 335) = 1.5, $p = .221$, but the interaction is significant, F (1,335) = 7.45, $p = .007$. This interaction effect describes the fact that the difference between epistemic

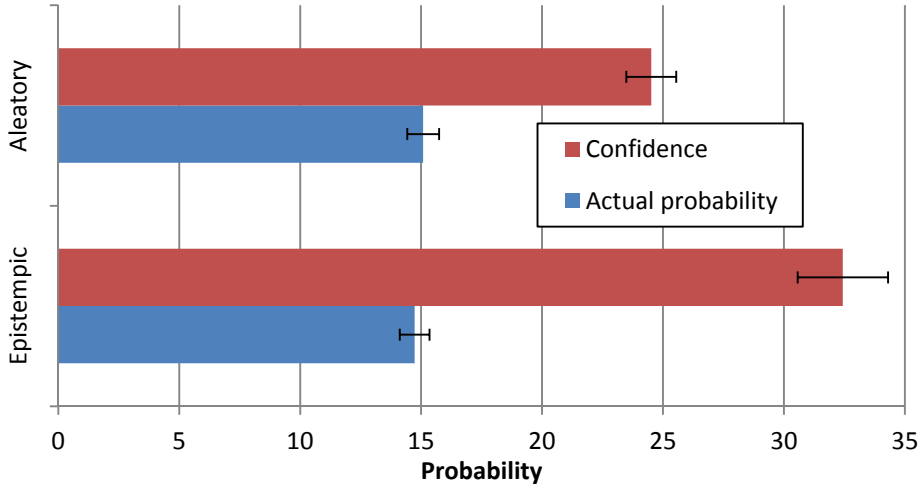
and aleatory frames was larger when the item-confidence judgment came after SPIES than when it came before. See Table 5.

Table 5. Mean Item-confidence (IC) and SDs of those means in the four different experimental conditions (Experiment 4).

	IC first	IC after SPIES
Epistemic framing	28.6 (23.5)	36.4 (23.7)
Aleatory framing	26.1 (12.0)	23.0 (15.0)

This greater confidence among those in the epistemic condition is not matched by greater accuracy, however. A 2 X 2 ANOVA on the distance between the actual most likely bin (Bin F) and the bin participants identified as being most likely reveals no significant effects of the experimental manipulations. The bin chosen as most likely was, on average, 1.64 bins away from Bin F for both epistemic and aleatory conditions. Consequently, those in the epistemic condition look dramatically more overconfident than do those in the aleatory condition. To test this, we entered both the estimated and actual probabilities as repeated measures into a mixed ANOVA, collapsing across IC/SPIES order, where the experimental manipulations served as between-subjects factors. The results reveal a main effect by which estimated probabilities (M = 28.32) systematically exceed actual probabilities (M = 14.92), F (1, 335) = 158.72, $p < .001$. This effect is qualified by an interaction by which those in the epistemic condition showed more overconfidence than did those in the aleatory condition, F (1, 335) = 15.09, $p < .001$. See Figure 5.

Figure 5. The actual and reported (item-confidence) probabilities of the ball landing in subjects' chosen bin, by experimental condition collapsing across IC/SPIES order (Experiment 4).



Contrary to our expectations, however, the manipulations did not affect SPIES responses. Those in the epistemic condition reported distributions with an average variance of 9.3, compared to 9.0 in the aleatory condition, a difference that is not statistically significant in a 2 X 2 ANOVA, F (1, 335) = 1.0, p = .32. Neither the main effect of order nor its interaction with epistemic vs. aleatory framing attained significance, F (1, 335) < 2.6, p > .1. As before, the participants were underprecise in the probability distributions they reported relative to the normative benchmark of a variance of 3.5. Compared to the actual variance in outcomes, 98% percent of participants reported probability distributions that were too wide, assigning too low a probability to the middle bin and probabilities that were far too high to the outer bins.

This failure to find an effect of the manipulation on the uncertainty participants reported via their SPIES distributions (participants in the aleatory and epistemic condition were equally underprecise by this measure) is remarkable, especially given two facts. The first is that the manipulation affected item-confidence judgments. This result highlights the fact that these two measures of confidence are not redundant with one another, nor is the psychology that drives

them. The second is that the manipulation could be criticized as being too strong because it confounds the nature of uncertainty (epistemic vs. aleatory) with the timing of the event (future vs. past). Prior research has found that describing risks as either epistemic (Fox & Ülkümen, 2011) or occurring in the past (Wright & Wishuda, 1982) contribute to increased certainty. The persistence of underprecision, replicating the result from Experiments 1-3, in the presence of both these manipulations suggests that persistent underprecision is not entirely explainable by either of these potential causes.

EXPERIMENT 5

Given the surprising persistence of underprecision when measured by confidence interval width and variance of subjective probability distribution, and the degree to which it is at odds with the prevailing consensus about the psychological causes of overconfidence in judgment, it is worth asking about the degree to which this result may be the product of the particular elicitations we have used. There is, of course, evidence suggesting that numerical probabilities might not be the most natural way for people to represent or express uncertainty (Brase, Cosmides, & Tooby, 1998; Gigerenzer, 1991). In order to address this concern, Experiment 5 employed a behavioral measure of uncertainty developed by Mamassian (2008) that parallels the way uncertainty should affect behavior in many everyday situations. We manipulated the rewards for under- vs. overestimating the right answer and observed whether this led people to shift their responses as much as they should have, normatively speaking.

Mannes and Moore (2013) found that people did not shift enough, consistent with excessive faith in the accuracy of their judgment. However, in their studies, participants made estimates about outcomes without known probability distributions. As a result, participants may

have appeared overprecise given their own (possibly faulty) beliefs. In the current study, we know the correct probability distribution for all outcomes and can specify what participants *should* believe and how they should shift their responses based on these correct beliefs. Study 5 tests whether we can replicate the overprecision implied by insufficient shifting found by Mannes and Moore (2013), while at the same time replicating the underprecision in reported beliefs illustrated so dramatically in Experiments 1-4.

Method

Participants were 67 MBA students enrolled in two sections of a decision making elective course at a west-coast business school in the United States. The sample size was determined by including all students present in class on the relevant day.

Procedure. As a part of a class exercise on understanding uncertainty, participants learned about the Quincunx and completed a SPIES elicitation that asked them to identify the probability the ball would land in each of the 11 bins.

They were then invited to bet on where the ball would land in each of 20 rounds. In the first round, half the participants read, "You must choose one of the eleven bins below by marking it with an X. If the ball lands in the bin you mark or one to the left, you will get 2 points. If the ball lands two or more bins to the left of the one you mark, you will get 1 point. If the ball lands to the right of the bin you mark, you will get 0 points for this round." Given these incentives and the actual probability distribution, the optimal shift was one bin to the right of the middle bin. The other half the participants had the flipped incentives (left/right), and their optimal shift was one bin to the left of the middle bin.

After everyone had chosen a bin, the instructor let a ball fall through the Quincunx displayed on the screen in the classroom so that everyone could see the realized outcome. Then

they placed their bets for Round 2, and the next ball was dropped. After 10 rounds, everyone's incentives flipped (left/right) for the remaining 10 rounds.

Results and Discussion

The results replicate the excessive breadth of the probability distributions reported with SPIES. On average, the distributions participants reported had a variance of 6.89 (SD = 3.3), which is significantly higher than the true variance of 3.5, $t(66) = 8.39$, $p < .001$, indicating underprecision. Fully 28% of participants reported roughly equal probabilities across all bins. The clear implication of these beliefs is that people should shift too much in response to the asymmetric incentives we gave them. To illustrate why, using the most dramatic example, if the probability distribution was *actually* flat, then the optimal shift would be 5 bins from the center (rather than the one bin implied by the actual binomial distribution).

And that is indeed what happened: participants shifted too much in response to the asymmetric incentives we gave them. On average, participants shifted their bets 1.82 bins (SD = 1.01 bins) from the middle, which is significantly further than the optimal shift of 1 bin, $t(66) = 6.66$, $p < .001$. Indeed, there is a positive correlation of .45 between the variance (width) of the SPIES distribution an individual reported and the amount they shift. There is also some evidence of learning over the 20 rounds: Shifting decreased from an average of 2.4 bins in the first round to 1.5 bins in the last.

However, participants' shifting is only excessive, reflecting underconfidence, conditional on accurate beliefs about the probability distribution. When judged relative to their *actual* beliefs, shifting is woefully insufficient, consistent with the results of Mannes and Moore (2013) and suggesting overconfidence. For each individual, it is easy to compute the degree of shifting that would have maximized expected value, given stated beliefs about the probability

distribution. For each possible amount of shifting, we simply computed its expected value given the probability distribution participants reported. For each of the 11 bins, we can compute its expected value given each participant's reported probability distribution. The one choice with the highest expected value was the best choice for that person, given the beliefs they reported. Conditional on these reported beliefs, participants should have shifted an average of 4.6 bins, when in fact they shifted an average of 2.4 bins in the first round, and these two are significantly different from each other by paired $t(66) = 9.12$, $p < .001$.

Experiment 5 reveals more evidence of underprecision in beliefs about chance devices with known probability distributions, using a behavioral measure (choosing how much to shift), not a reported probability. At the same time, we also replicate prior results suggesting overprecision, as implied by insufficient shifting conditional on participants' own beliefs.

GENERAL DISCUSSION

The results presented here offer a surprise. Although we successfully replicate overprecision for Experiments 1-5, we also simultaneously find what appears to be underprecision. While a few results have found what might be underprecision in forecasts of time series data (see Lawrence et al., 2006, for a useful review), those studying overprecision in judgment have routinely assumed that low hit rates (implying overprecision) *result* from overly narrow subjective probability distributions (Alpert & Raiffa, 1982; Soll & Klayman, 2004; Tversky & Kahneman, 1974). Our data contradict this assumption. Instead, the probability distributions our otherwise overly precise participants report are far too wide. We were only able to employ this test using chance devices of known uncertainty. The results suggest the possibility that, if only we knew the true error distributions around people's beliefs for other

contexts of judgment, we might again observe that reported distributions would be too wide instead of being too narrow.

Reconciling contradictory results

The apparently contradictory facts that confidence intervals were too wide but include the correct answer too rarely can be reconciled by the fact that participants' answers are so inaccurate. The 90% confidence intervals are too wide, but they are centered so far from the correct answer that hit rates are still below 90%. Does that not imply that they are, at least relative to the accuracy standard of examining actual hit rates, too narrow? Answering this question raises profound issues of what it means to know something and how we come to be sure of our knowledge.

Most of the time, or at least in most prior studies of overprecision, people are operating in domains where uncertainty is more epistemic than aleatory. Under these circumstances, it is not possible to compute an error distribution for each individual, since that requires knowing what that person knows. In formulating their answers, people must rely on their own internal guesses about how wrong they might be when estimating a confidence interval around a best guess. In order to do so, they must appreciate all the ways in which their original estimate could be wrong and, crucially, what else could be right instead.

This may be asking too much. People will believe something, and they will not always be right. We cannot also assume that people are aware of all the knowledge that they lack. The one thing we can count on is that there will be variability in what people know and what they believe, even when they are all forecasting the same Quincunx machine. These differences, however, are between people. Within an individual, each one of us attempts to form the most

accurate and insightful beliefs we can. We believe what we believe because we believe it to be true (Gilbert, 1991; Schulz, 2010).

Can our results help us understand overprecision in judgment?

We began by noting that overprecision is both the most robust and the least understood form of overconfidence. Can the results we present offer insight into why it happens? We see our results offering both negative and positive conclusions. On the negative side, we believe our results speak against some explanations that researchers have offered for overprecision.

Anchoring. Tversky and Kahneman (1974) describe overly narrow confidence intervals in judgment as an instance of anchoring and insufficient adjustment. In this view, participants create confidence intervals that are too narrow because they focus on their best estimate of an outcome then insufficiently adjust up and down from this value when creating their confidence interval bounds. Two features of our results are at odds with anchoring. First, we find confidence intervals that are too wide when we can specify how wide they should be. Following the reasoning of Tversky and Kahneman, this would suggest excessive adjustment from participants' best estimate anchor.

Second, helping set the anchor by first completing a best guess such as an item-confidence judgment ought to increase precision of a subsequent SPIES judgment. We do not find that it does. Specifically, in Experiment 4, participants reported probability distributions that were (equally) excessively wide in both the epistemic and aleatory conditions and whether they completed a best guess (item-confidence item) before or after reporting their full subjective probability distribution. There was also no significant interaction between these manipulations. Anchoring on a best guess and insufficiently adjusting from this best guess cannot explain our findings, then, of simultaneous over- and underprecision.

Conversational norms. The notion that conversational norms favor the expression of greater precision over greater accuracy is an appealing explanation for overprecision (Yaniv & Foster, 1995). Using this reasoning, the precision or "graininess" of a judgment is evaluated in terms of its informativeness and accuracy. While wider confidence intervals are likely more accurate, they are less informative because they give a wide range of possible outcomes. Narrower confidence intervals, on the other hand, are more informative because they restrict the range of possible outcomes but can come at the cost of accuracy by doing so.

Yaniv and Foster (1995) find that individuals receiving judgments often prefer informativeness over accuracy. If people making judgments are aware of this and/or hold this preference for informativeness themselves, it would make sense why individuals would report overly narrow confidence interval rates. These confidence intervals would serve a functional purpose of providing more information despite limiting accuracy. However, this functional account represents a poor explanation for our results, given that our participants gave us confidence intervals that were so much wider than they should have been.

Naïve intuitive statistician. This explanation rests on the notion that when estimating an unknown quantity, people accurately sample information from their previous experience, but fail to distinguish how representative (or not) the samples are (Juslin et al., 2007). Similar to the anchoring argument above, the naïve intuitive statistician explanation suggests that when individuals have previous knowledge or experience with an event, they do not use this sample information appropriately to make judgments of future events. Rather than using sample properties to estimate properties of the population of all possible outcomes, individuals instead anchor on sample properties and use this information *as if it were* the full range of possible events, often leading to too narrow precision around estimates.

In Experiment 3, we varied the amount of experience participants had with the ball drop machine. In a sense, we varied the number of sample outcomes participants had exposure to, particularly in the slow and fast machine conditions. The naïve intuitive statistician account would predict that participants would give fairly reasonable best estimates of a future ball drop outcome based on their previous experience with the machine, yet would also overly rely on these same sample outcomes when providing the certainty around this estimate and provide overly narrow subjective probability distributions due to this overreliance. However, our results appear to be at odds with this account, both because participants' best estimates are so far off (even in the fast machine condition, only 64% of participants choose the correct bin) and because their resulting probability distributions are so wide.

Practical Implications

Noting the widespread prevalence of overprecision in judgment, scholars have recommended that organizations intervene to correct for human bias (Camerer, Issacharoff, Loewenstein, O'Donoghue, & Rabin, 2003; Heath, Larrick, & Klayman, 1998; Thaler & Sunstein, 2008). If structural engineers' margins of safety are too narrow, their firms can multiply their computations by some "safety factor." If software engineers' estimates for how long it will take to complete development projects is systematically biased, their organizations can adjust their estimated completion times before deciding what to tell the customer about when to expect a finished product.

We believe our results recommend caution when devising such organizational repairs to human bias. Unless we are sure which way the bias goes, it is difficult to design a single all-purpose debiasing intervention. Prior research would have made it easy to conclude that overprecision in judgment is so rampant that wise organizations ought to be correcting for biases

in human judgment by widening confidence intervals. However, if people's confidence intervals are actually too wide, then widening them even more could create more problems than it solves.

Conclusions and Future Directions

On the more positive side, we believe our results paint a new picture of the epistemic and psychological forces at work behind overprecision in judgment. Despite their excessive width, people look overprecise by traditional measures because their estimates are so bad. This result strongly suggests that overprecision is attributable to people's failure to accurately calibrate the error distributions around their own beliefs. But this conclusion raises more questions. What, for instance, is the phenomenological experience of overprecision or underprecision in judgment? When I report a confidence interval that is many times as wide as it should be, do I feel unsure of myself? What sort of self-assurance accompanies a person's claim that they know where the ball will land in the Quincunx machine? Do these feelings predict actions such as a willingness to bet on their beliefs? We hope future research will answer these questions.

We set out to test one of the basic assumptions that researchers have made when they have attempted to explain results that imply overprecision in judgment: that the subjective probability distributions from which people derive their confidence judgments are too narrow. When assessed against accuracy in hit rates using traditional approaches, they do indeed appear to be too narrow. However, when we test them against normative benchmarks for the width of probability distributions, they appear to be far too wide. We also learn that while subjective probability distributions become narrower with experience, they still typically remain too broad by normative standards. By using chance devices that have a known probability distribution, we are able to discern the difference between these two types of judgment and find that overly narrow subjective probability distributions cannot, at least in our studies, cause the overprecision

that we find. We believe that these surprising results call into question the very definition of what it means to be overconfident in one's judgment, and highlight a need for future research into the psychological causes of both over- and under-precision.

References

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes*, *49*(2), 188–207.

Brase, G. L., Cosmides, L., & Tooby, J. (1998). Individuation, counting, and statistical inference: The role of frequency and whole-object representations in judgment under uncertainty. *Journal of Experimental Psychology: General*, *127*(1), 3–21.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Camerer, C. F., Issacharoff, S., Loewenstein, G., O'Donoghue, T., & Rabin, M. (2003). Regulation for Conservatives: Behavioral Economics and the Case for "Asymmetric Paternalism." *University of Pennyslvania Law Review*, *151*(3), 1211–1254.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107*(4), 659–676.

Fox, C. R., & Ülkümen, G. (2011). Distinguishing two dimensions of uncertainty. In W. Brun, G. Keren, G. Kirkebøen, & H. Montgomery (Eds.), *Perspectives on Thinking, Judging, and Decision Making* (pp. 21–35). Oslo: Universitetsforlaget.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases." *European Review of Social Psychology*, *2*(1), 83–115.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist*, *46*(2), 107–119.

Gino, F., & Moore, D. A. (2007). Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making*, *20*(1), 21–35.

Haran, U., Moore, D. A., & Morewedge, C. K. (2010). A simple remedy for overprecision in judgment. *Judgment and Decision Making*, *5*(7), 467–476.

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Sciences*, *1*(2), 78–82.

Heath, C., Larrick, R. P., & Klayman, J. (1998). Cognitive repairs: How organizational practices can compensate for individual shortcomings. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior, Vol. 20: An annual series of analytical essays and critical reviews.* (pp. 1–37). Greenwich, CT, USA: Jai Press, Inc.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*(8), 534–539.

Hertwig, R., & Erev, I. (2009). The decision-experience gap in risky choice. *Trends in Cognitive Sciences, 13*(12), 517–523.

Jain, K., Mukherjee, K., Bearden, J. N., & Gaba, A. (2013). Unpacking the Future: A Nudge Toward Wider Subjective Confidence Intervals. *Management Science*. doi:10.1287/mnsc.1120.1696

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review, 114*(3), 678–703.

Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting, 22*(3), 493–518. Retrieved from http://www.sciencedirect.com/science/article/pii/S0169207006000501

Mamassian, P. (2008). Overconfidence in an objective anticipatory motor task. *Psychological Science, 19*(6), 601–606.

Mannes, A. E., & Moore, D. A. (2013). A behavioral demonstration of overconfidence in judgment. *Psychological Science, 24*(7), 1190–1197.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2), 502–517.

Moore, D. A., Tenney, E. R., & Haran, U. (2015). Overprecision in judgment. In G. Wu & G. Keren (Eds.), *Handbook of Judgment and Decision Making*. New York: Wiley.

Plous, S. (1995). A comparison of strategies for reducing interval overconfidence in group judgments. *Journal of Applied Psychology, 80*(4), 443–454.

Schulz, K. (2010). *Being wrong*. New York: Ecco.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics, 1*(1), 43–61.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(2), 299–314.

Tannenbaum, D., Fox, C. R., & Ülkümen, G. (2014). Judgment extremity and accuracy under epistemic versus aleatory uncertainty. *Unpublished Manuscript*.

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Yale University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–31. doi:10.1126/science.185.4157.1124

Wright, G., & Wishuda, A. (1982). Distribution of probability assessments for almanac and future event questions. *Scandinavian Journal of Psychology*, *23*, 219–224.

Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General*, *124*(4), 424–32.